Berkeley
UNIVERSITY OF CALIFORNIA

CHAN ZUCKERBERG
BIOHUB

Veridical Data Science for Precision Medicine: subgroup discovery through staDISC

Bin Yu

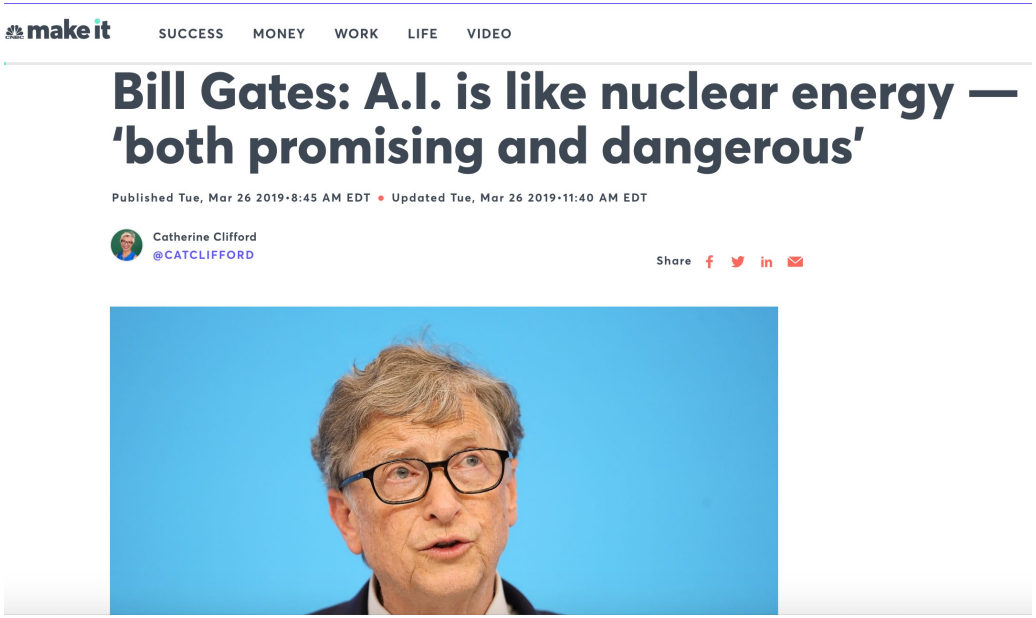
Statistics and EECS, UC Berkeley

"Bringing Artificial Intelligence to the Bedside"

Purdue University

April 23, 2021

2021 AI is part of modern life

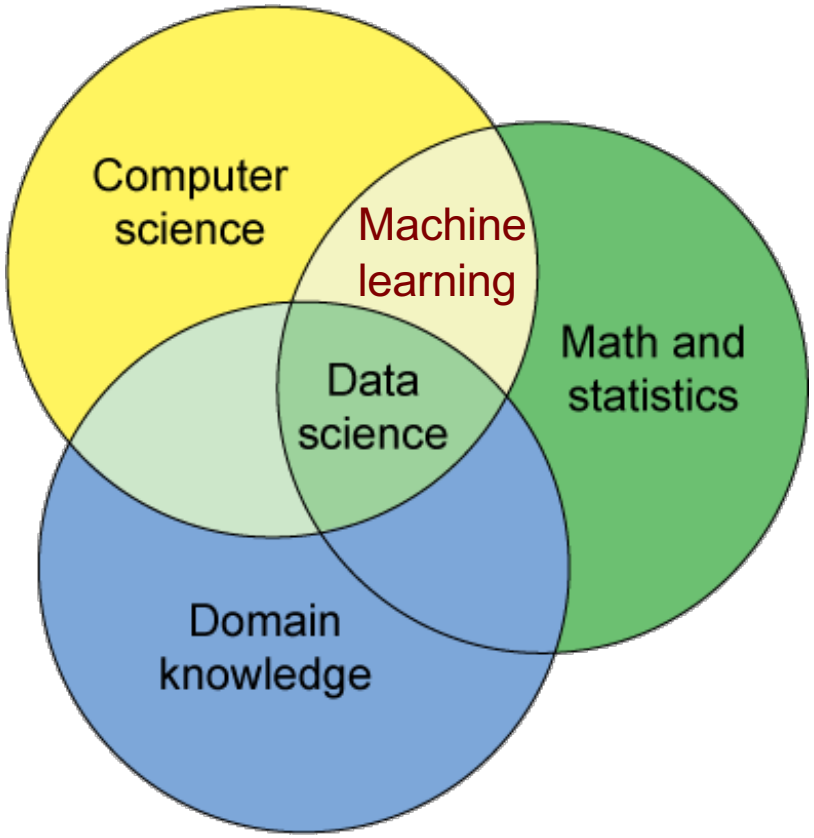


2019

Alexa, Siri, …
Wearable health devices
Streaming videos, on-line gaming, …
On-line news
Self-driving cars
Election campaigns
Precision medicine
Biology
Neuroscience
Cosmology
Material science
Chemistry
Law
Political science
Economics
Sociology
…

Data science is a key element of AI

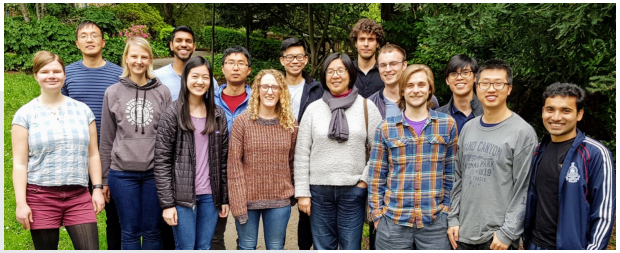
Conway's Venn Diagram

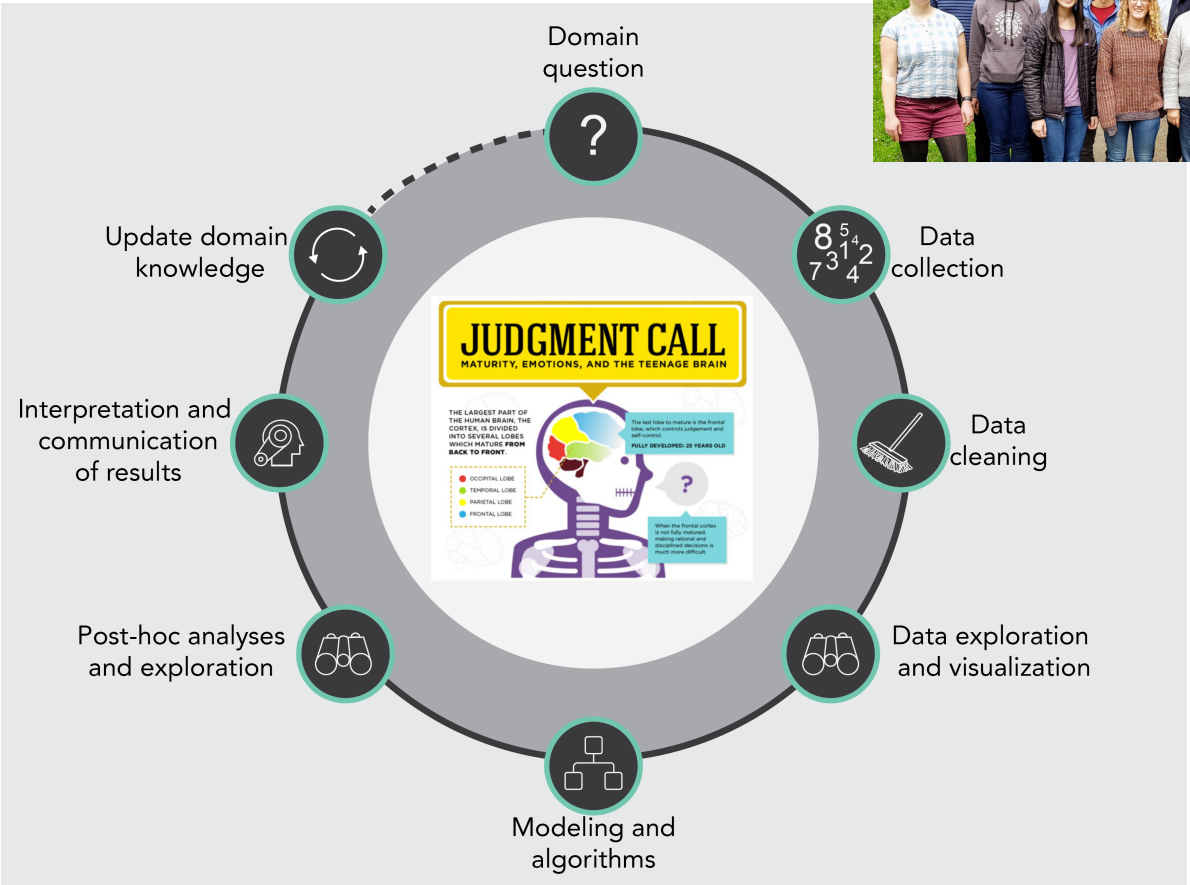


Goal:

combine data with domain knowledge to make decisions and generate new knowledge

Data Science: a process or a life cycle





Missing: quality control and standardization of the process

Image credits: R. Barter and toronto4kids.com

Trustworthy AI:
two complementary approaches

- **Best practices to maximize the promise (preventative)**

- Risk management to reduce the danger (intervention)

Veridical Data Science

Extracts reliable and reproducible information from data,
with an enriched technical language to communicate
and evaluate empirical evidence in the context of
human decisions and domain knowledge

It realizes promises and mitigates dangers of AI.

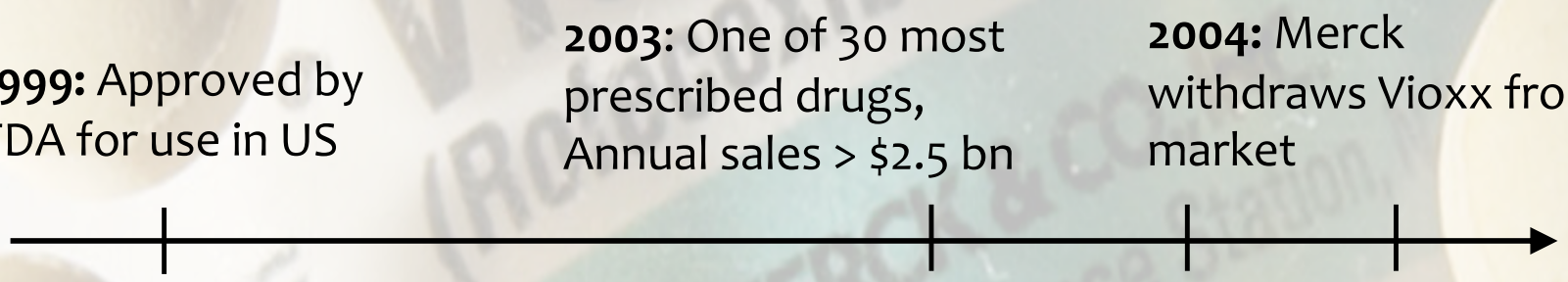
Precision Medicine

a problem of medicine and data science


VIOXX® 25 mg
(Rofecoxib Tablets)
MERCK & CO., INC.
Whitehouse Station,
Vioxx

- Regular use of non-steroidal anti-inflammatory drugs (NSAIDs) increases risk of gastro-intestinal perforations, ulcers and bleeding

- Vioxx is a *selective* NSAID that was demonstrated to have lower increased risk compared to non-selective NSAIDs



1999: Approved by FDA for use in US

2003: One of 30 most prescribed drugs, Annual sales > $2.5 bn

2004: Merck withdraws Vioxx from market

2001-2004: Study found that Vioxx **increased the risk of thrombotic cardiovascular events**

2005: FDA says that benefits may outweigh risks, may return to market

Vioxx problem as a data science problem

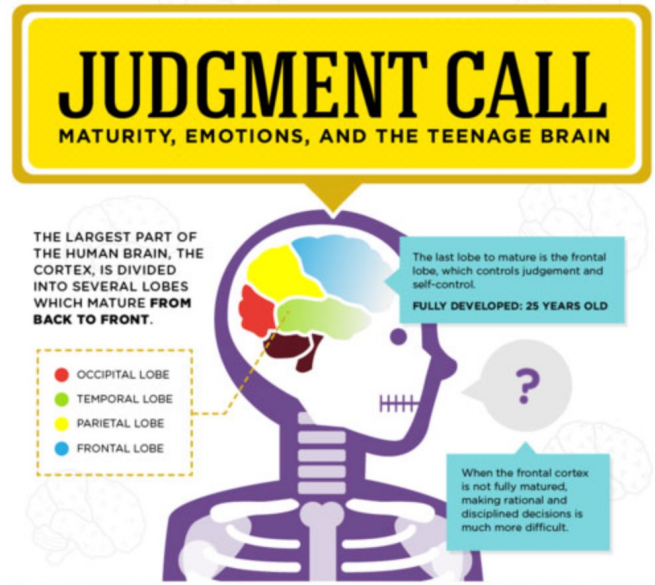
- Medical question: are there subgroups of people who only benefit from Vioxx?

- What data to use? Cleaned?

- Look at data: summaries, data plots, …

- Modeling: heterogeneous treatment effect estimation

- Interpretation of data results



- Validation: is Vioxx only beneficial to future patients in these subgroups?

PCS framework for veridical data science and trustworthy AI

PCS framework Yu and Kumbier (PNAS, 2020)

Three principles of data science : PCS

Predictability (**P**) (ML and Stats)

Computability (**C**) (ML)

Stability (**S**) (Stats)

It unifies, streamlines and expands on ideas and best practices in ML and Stats

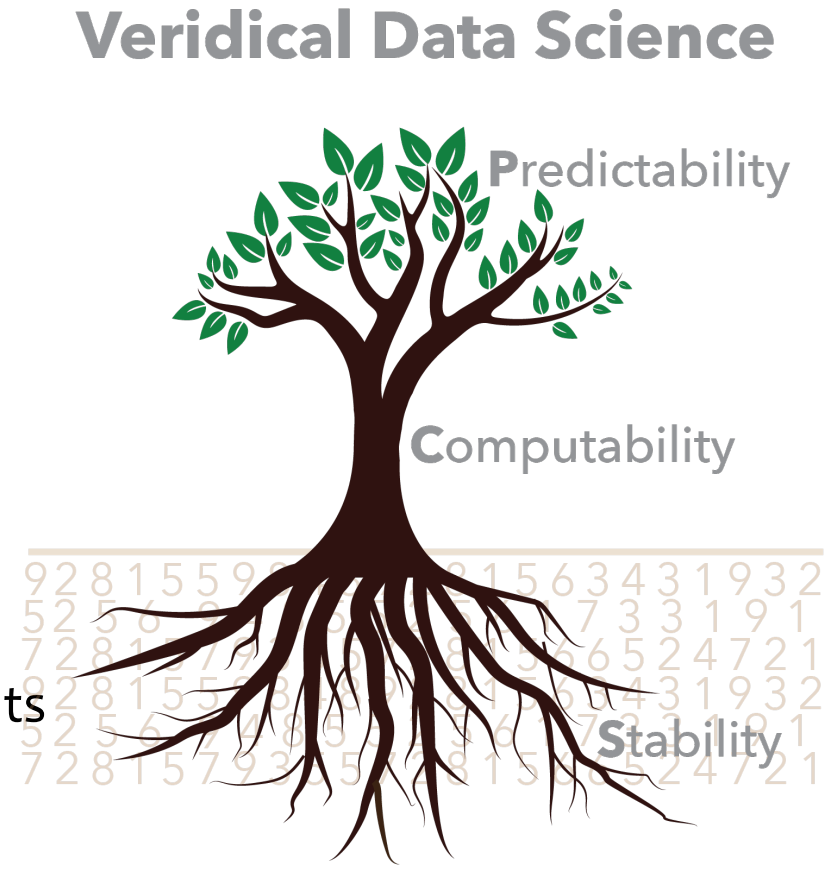


Image credit: R. Barter

The stability principle

Reproducibility is imperative for any scientific discovery. More often than not, modern scientific findings rely on statistical analysis of high-dimensional data. At a minimum, reproducibility manifests itself in ***stability*** *of statistical results relative to* ***reasonable perturbations*** *to data and to the model used.*

- Yu (2013) [Stability]

PCS in a nutshell

Predictability for reality check (in terms of a user defined prediction error)
Stability analysis over human choices including data and model perturbations
Computability is implicit in P and S steps.

data science life cycle

Intuitively, stability analysis is about shaking every part so it doesn't break.

Precisely, it assesses impact of a reasonable "perturbation" defined by the user on a stability metric (e.g based on medical knowledge) also defined by the user – broad and flexible

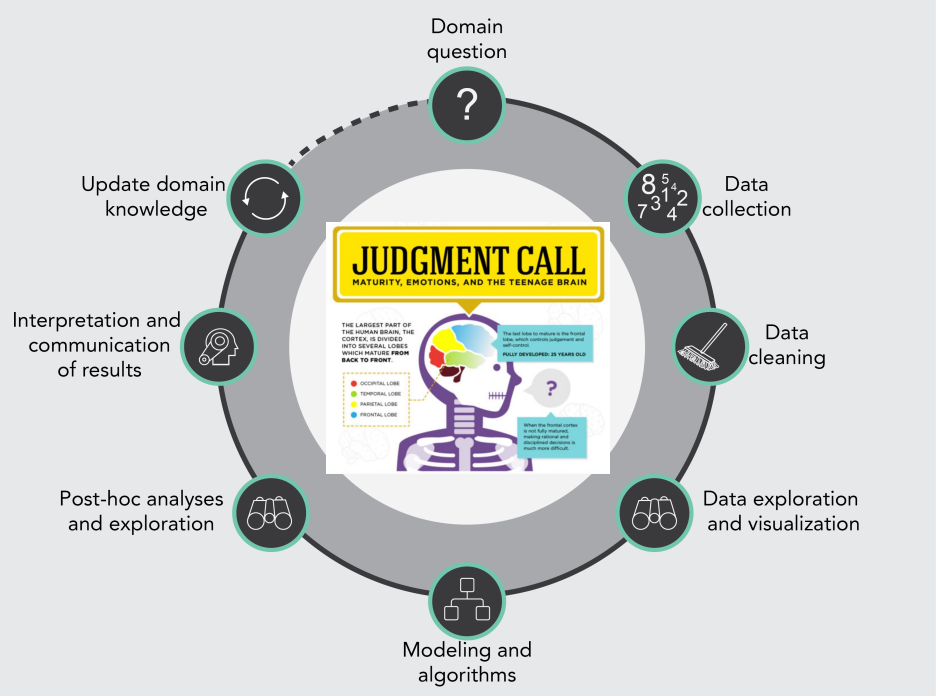
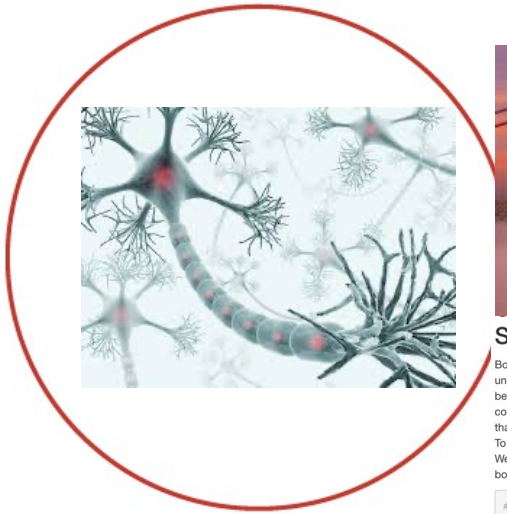


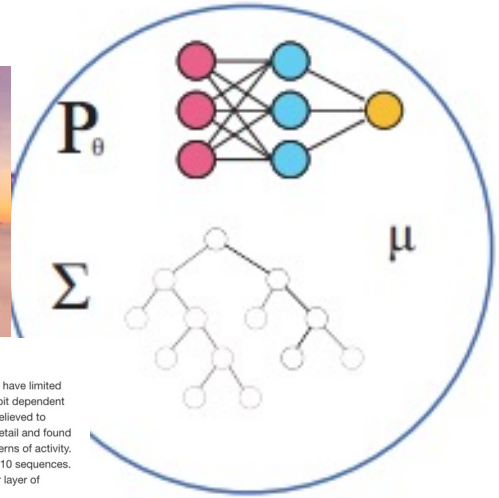
Image credits: R. Barter and toronto4kids.com

PCS documentation on github (JupyterNotebook) bridges reality and models

Reality

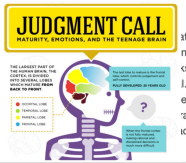


Models



Stability formulation

Bootstrap sampling is a widely accepted perturbation
understanding of the dependencies. However, sequen
behavior that is possible to account for. In particular, e
confer robustness to regulatory processes (Hong, Hen
that over 70% of loci they examined have anywhere fr
To account for this potential dependency along the ge
We define the stability of an interaction to be the propo
bootstrap samples using the 3 proposed perturbation



at is a useful baseline for data where we have limited
me space (i.e. nearby on the DNA) exhibit dependent
ks known as "shadow enhancers" are believed to
l. 2016) studied shadow enhancers in detail and found
et al. 2016) with highly overlapping patterns of activity.
ap perturbations using blocks of 5 and 10 sequences.
across $B = 100$ RFs trained on an outer layer of

```
# Block bootstrap for blocks of size 5 and 10
block5.tr <- makeBlocks(gene.coords, idcs=train.id, size=5)
block10.tr <- makeBlocks(gene.coords, idcs=train.id, size=10)
block5.tst <- makeBlocks(gene.coords, idcs=test.id, size=5)
block10.tst <- makeBlocks(gene.coords, idcs=test.id, size=10)
```

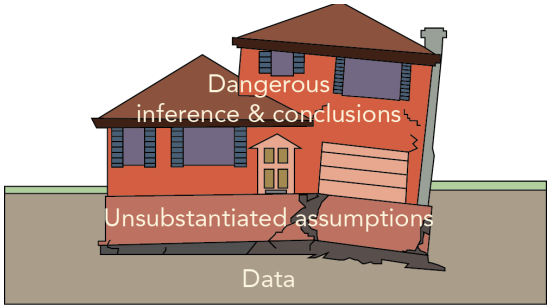


Image credit: Rebecca Barter

What is PCS to a doctor?

"The PCS framework builds a working relationship between data and the clinical world."

"PCS is a `look under the hood' to ensure that the conclusions found are what the data genuinely suggest. In all, PCS is a holistic approach to helping the clinician understand, interpret, and build the science we need to help our patients."

"Looking under the hood"

Which one do we want to see?

Image credits: https://www.csi.state.co.us/cars-looking-under-the-hood/
https://medium.com/@dottykeenan/looking-under-the-hood-of-your-agile-delivery-3f2f4ba7b5f

Dr. Aaron Kornblith, ER, UCSF

main medical collaborator on

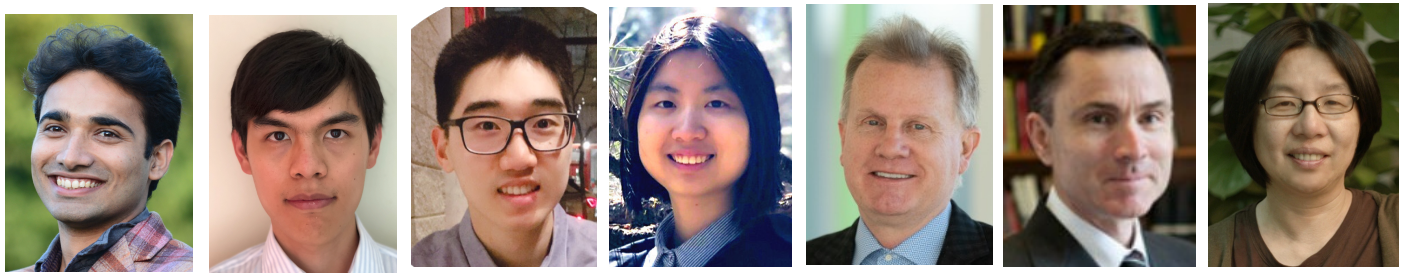
PCS stress testing of PECARN CDR

StaDISC:
Stable Discovery of Interpretable Subgroups via Calibration

A PCS case study in the hope of developing new clinical decision rule (CDR)

A collaborative project

- Raaz Dwivedi, Yan Shuo Tan, Briton Park, Mian Wei, Kevin Horgan, David Madigan, Bin Yu



Stable discovery of interpretable subgroups via calibration in causal studies.
International Statistical Review, 2020
also at arXiv:2008.10109

Q: Can we find a subgroup of patients who benefit from **Vioxx** but do not suffer from its drawbacks?

We want subgroups to be predictive, stable, and interpretable, or we want a CDR.

Judgment calls ubiquitous in data science life cycle

- **Which problem to work on**
- Which data sets to use
- How to clean
- What plots
- What data perturbations
- What algorithm perturbations
- What post-hoc plots/results
- What interpretations
- What conclusions

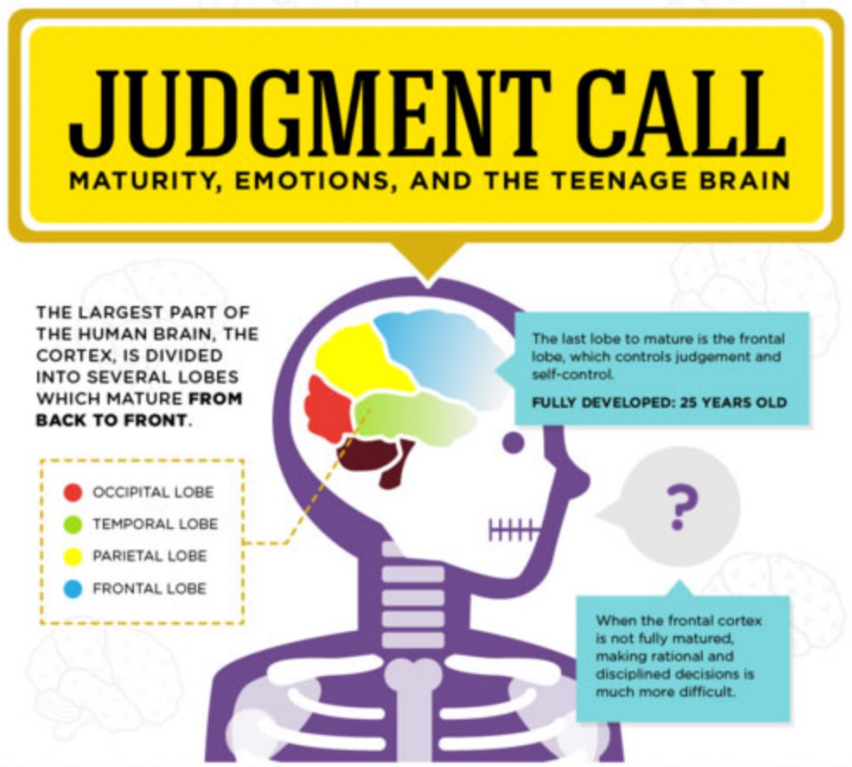


Image credits: toronto4kids.com

The VIGOR study: Vioxx GI Outcomes Research

- 1999-2000 **randomized controlled trial** by Merck with **8076 patients** who had rheumatoid arthritis
- Conducted at 301 centers in 22 countries
- Treatment arm: **Vioxx** vs Control arm: **Naproxen**

Bombardier et al.. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *New England Journal of Medicine*, 343(21):1520–1528, 2000

The VIGOR study: Vioxx GI Outcomes Research

Treatment arm: **Vioxx** vs Control arm: **Naproxen**

Outcome	ATE	Control Arm Rate
Gastro-intestinal (GI) event	-1.6%	3.0%
Thrombotic cardiovascular (CVT) event	0.6%	0.4%

Bombardier et al.. Comparison of upper gastrointestinal toxicity of rofecoxib and naproxen in patients with rheumatoid arthritis. VIGOR Study Group. *New England Journal of Medicine*, 343(21):1520–1528, 2000

Neyman-Rubin framework [Rubin '74]

Randomized Controlled Trial (RCT) is a special case

- Assume a superpopulation: data units can be viewed as random samples from a population of patients who might benefit from Vioxx

- Randomized experiment guarantees that

$Y_i(T_i), X_i | T_i = a$ has same distribution as $(Y_i(a), X_i)$ for $a = 0,1$

Neyman-Rubin framework [Rubin '74]

- Average Treatment Effect (ATE):

$$\tau_{ATE} \coloneqq \mathbb{E}_{\mathbb{P}}[Y_i(1) - Y_i(0)]$$

- Conditional Average Treatment Effect (CATE):

$$\tau(x) \coloneqq \mathbb{E}[Y_i(1) - Y_i(0) | X = x]$$

e.g. average effect for people of age x= 65

- Subgroup **CATE**: Given a subgroup $\mathcal{G} \subset \mathcal{X}$
  - $\tau_{\mathcal{G}} \coloneqq \mathbb{E}[Y_i(1) - Y_i(0) | X \in \mathcal{G}] = \mathbb{E}[\tau(X) | X \in \mathcal{G}]$

Translating medical goal (CDR) to a data science problem

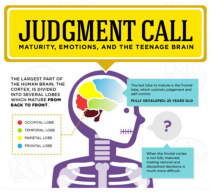
Find **interpretable** $\mathcal{G}$ for which $\tau_{\mathcal{G}}$ is **smaller** than τ_{ATE}.

Feature Engineering
causes data perturbations:

different people create features
in different ways

Feature engineering through threshold choice (stability analysis later)

16 binary features (covariates) including 6 features binarized by us



- Demographics (5):
  - Gender
  - Race, country, elderly, obese (*binarized*)

- Lifestyle risk factors (2):
  - Smoking, drinking (*binarized*)

- Medical risk factors (9):
  - (6) Medical history (GI, diabetes, hypertension, hypercholesterolemia, atherosclerosis, FDA indication for asiprin)
  - (3) Prior use of other medication (steroids, NSAIDS, NAPRXN)

Data Split
causes data perturbations:

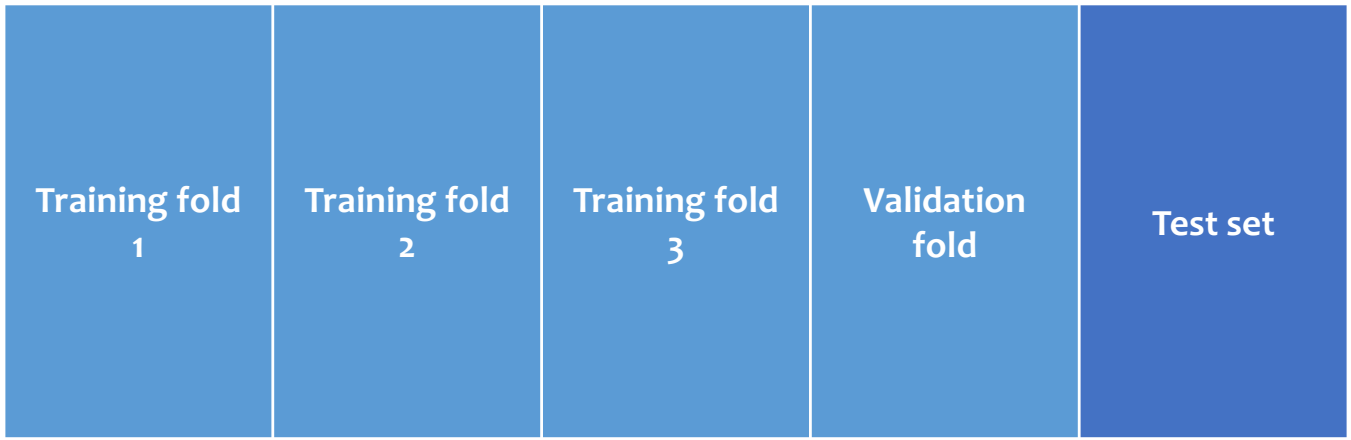
different people create splits
in different ways

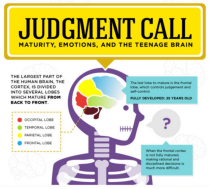
Data perturbations to assess P

Data splitting (stratified by treatment & outcome)

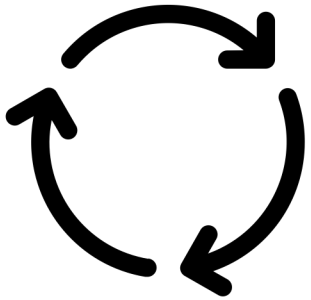
Training folds (cross-validation)

Works well if data units are symmetric





Surrogates for future data



12 settings (or perturbations):
permute val. fold 4 times * re-split 3 times

Data perturbations (recent)

- **Adversarial attacks to deep learning algorithms – stress test**

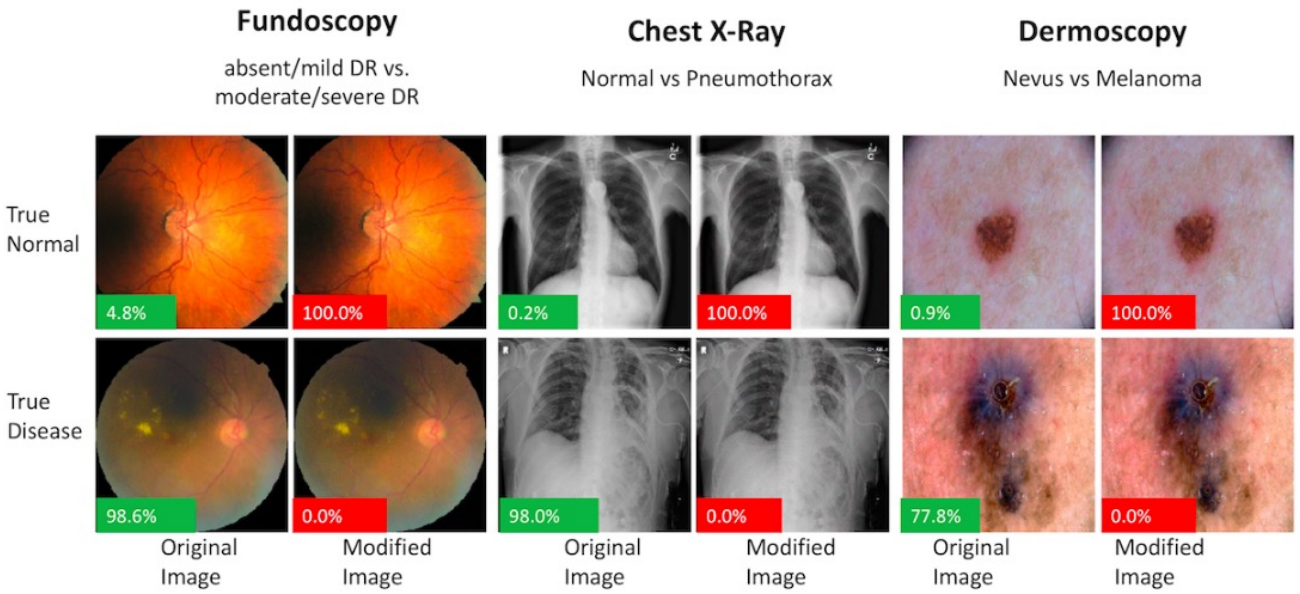


Image credits:groundai.com

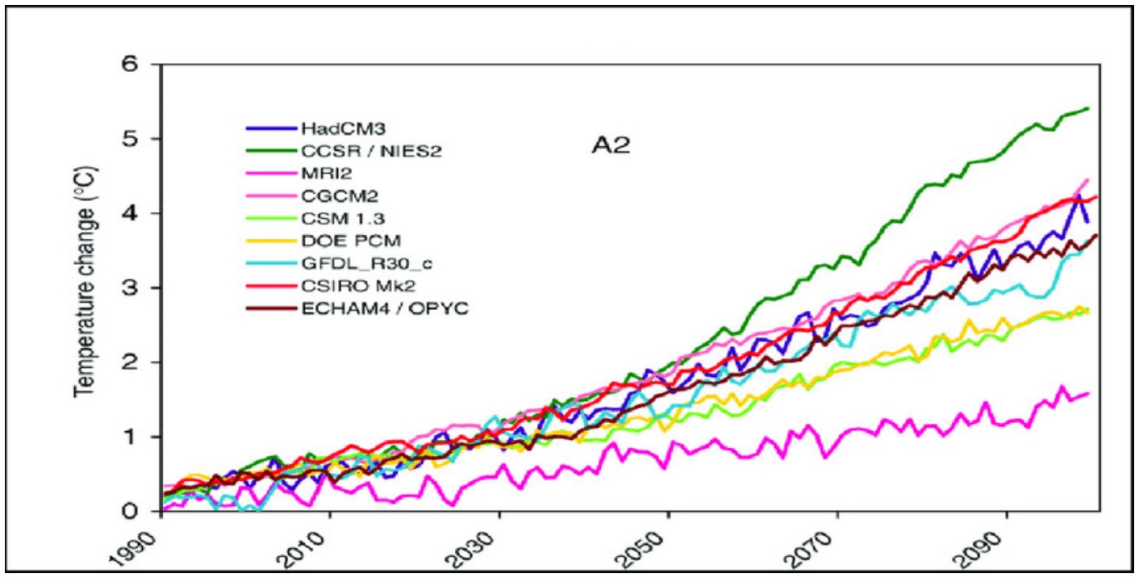
Method/Algorithm Choice
causes model perturbations:

different people prefer different methods/algorithms

Model/algorithm perturbations (new)

- Researcher to researcher (or team to team) perturbation

9 climate models



The change in global-mean temperature estimated by nine climate models forced by the SRES A2 emission scenario. (Source: IPCC TAR, Chapter 9)

Global mean-temp change

Two routes when facing so many data/model choices

- Cherry-picking: commonly used in published papers

Find the best feature engineering, best data split, and best model using training set –

- PCS:

P+S: use prediction error on validation set to screen to a filtered model set by averaging performance cross different data perturbations

S: use stability analysis to "combine" in the filtered set --
"average", "majority voting", "ensemble", …

Image credit:https://www.leaprate.com/financial-services/charges/sec-uses-data-analysis-to-detect-cherry-picking-by-broker/

Recall CATE

Conditional Average Treatment Effect (CATE):

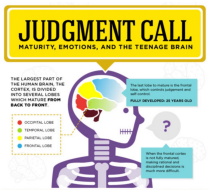
$$\tau(x) \coloneqq \mathbb{E}[Y_i(1) - Y_i(0) | X = x]$$

e.g. Subgroup average treatment effect for people of age x=65

Model perturbations: CATE

- Heterogeneous effect estimation via CATE function $\hat{\tau}(x)$ from samples, and use $\hat{\tau}(x)$ to identify subgroups

- Many non-parametric CATE models (we used 17 of them)



  - Tree-based methods: Use causal split condition

    - Causal tree [Athey-Imbens '16]
    - Causal forest [Wager-Athey '18]
    - BART and BCF [Chipman-George-McCulloch '10, Hill '12, Hahn-Murray-Carvalho '20]

  - Meta-learner framework: Solve regression sub-problems using ML

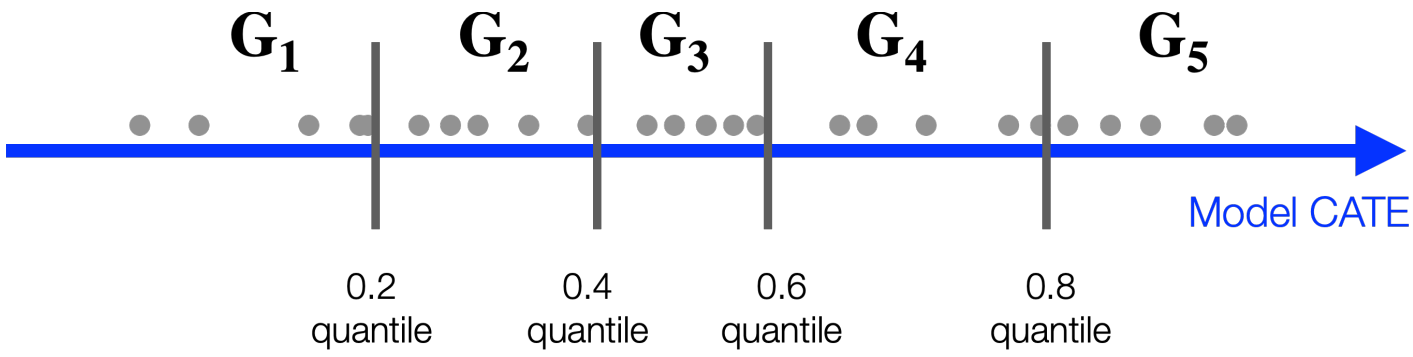
    - S-learner: *Y(T) ~ (X,T)*
      [Hill 2011, Green-Kern 2012]
    - T-learner: *Y(1) ~ (X,T) and Y(0) ~ (X,T)*
      [Foster-Taylor-Ruberg'11, Imai-Ratkovic '13, …]
    - **X-learner** [Kunzel et al. '19]
    - R-learner [Nie-Wager '20]

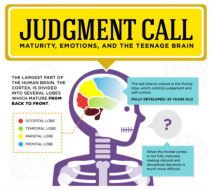
Prediction screening of 17 models via calibration and t-score

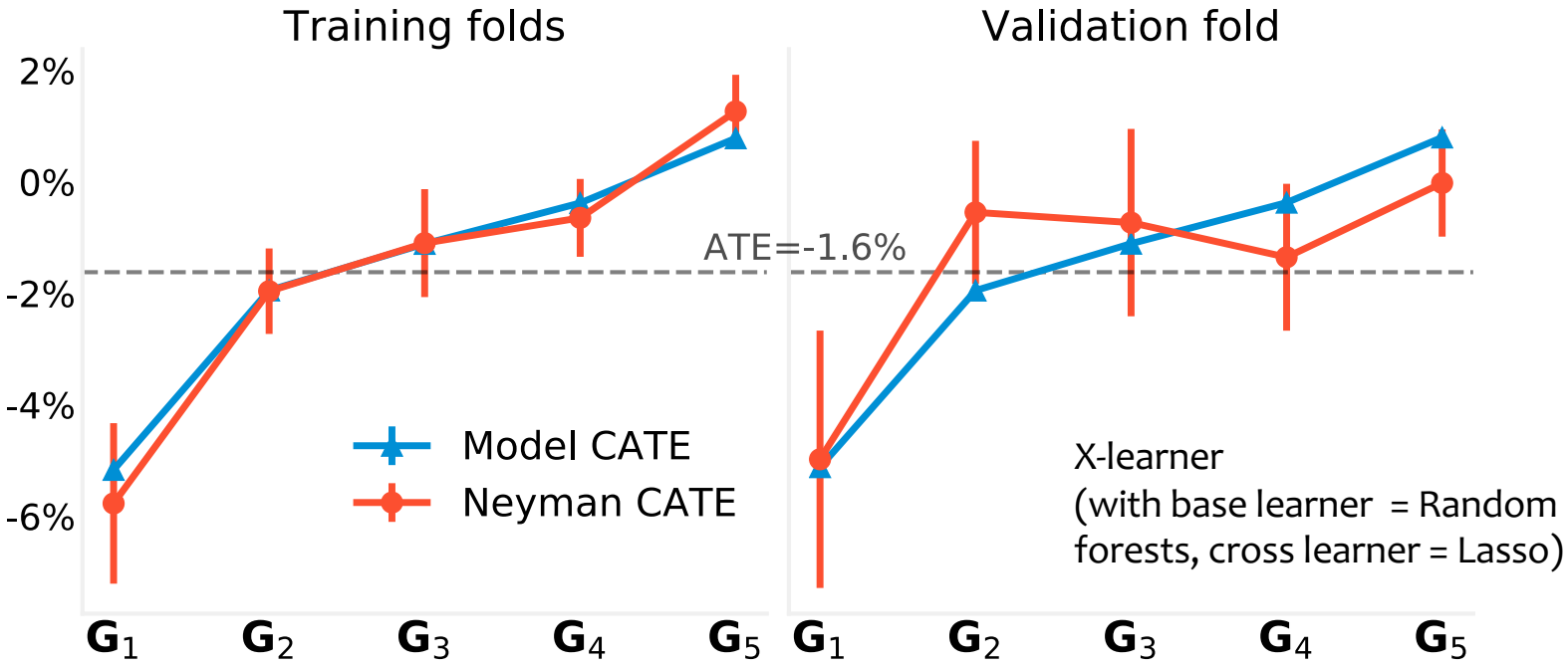
averaged over different data perturbations

Prediction-screening via calibration for CATE

Visual Assessment for sub groups indexed by G







Prediction screening via calibration: *Take-aways (for VIGOR study)*

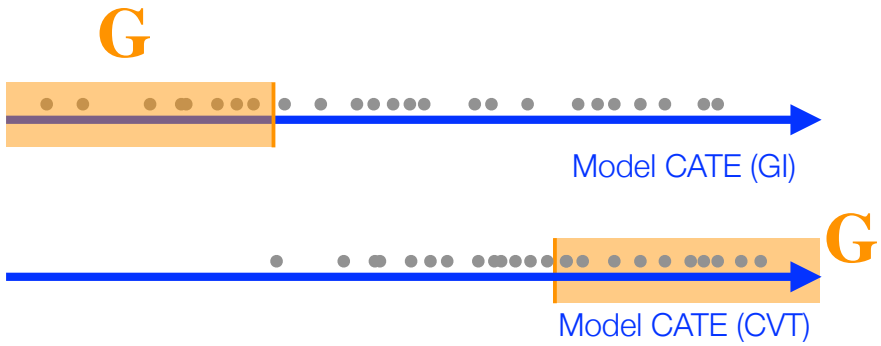
- CATE models do not have "good generalization" on the whole dataset
- Bottom quantile-based subgroups (for GI event) and top quantile-based subgroups (for CVT event) are relatively more *stable and across 17 models*
- *But such subgroups are different for each CATE model*

Next steps

- Which CATE models to use to identify subgroups
- How to turn subgroups into clinically interpretable subgroups (or CDR)

t-score
measures gain of subgroup over ATE

- Quantile-based subgroups

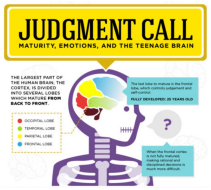


- Standardize subgroup CATE (t-statistics)

$$\mathbb{T}_{\mathbf{G}} := \frac{\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\mathrm{ATE}}}{\sqrt{\widehat{\mathrm{Var}}(\widehat{\tau}_{\mathbf{G}} - \widehat{\tau}_{\mathrm{ATE}})}}$$

- For each **model+data split pair** $(M, \mathfrak{D})$, compute the avg. t-statistics across folds and different (overlapping) quantile groups

Stability analysis with t-score



7 t-scores for each of 17 CATE models from

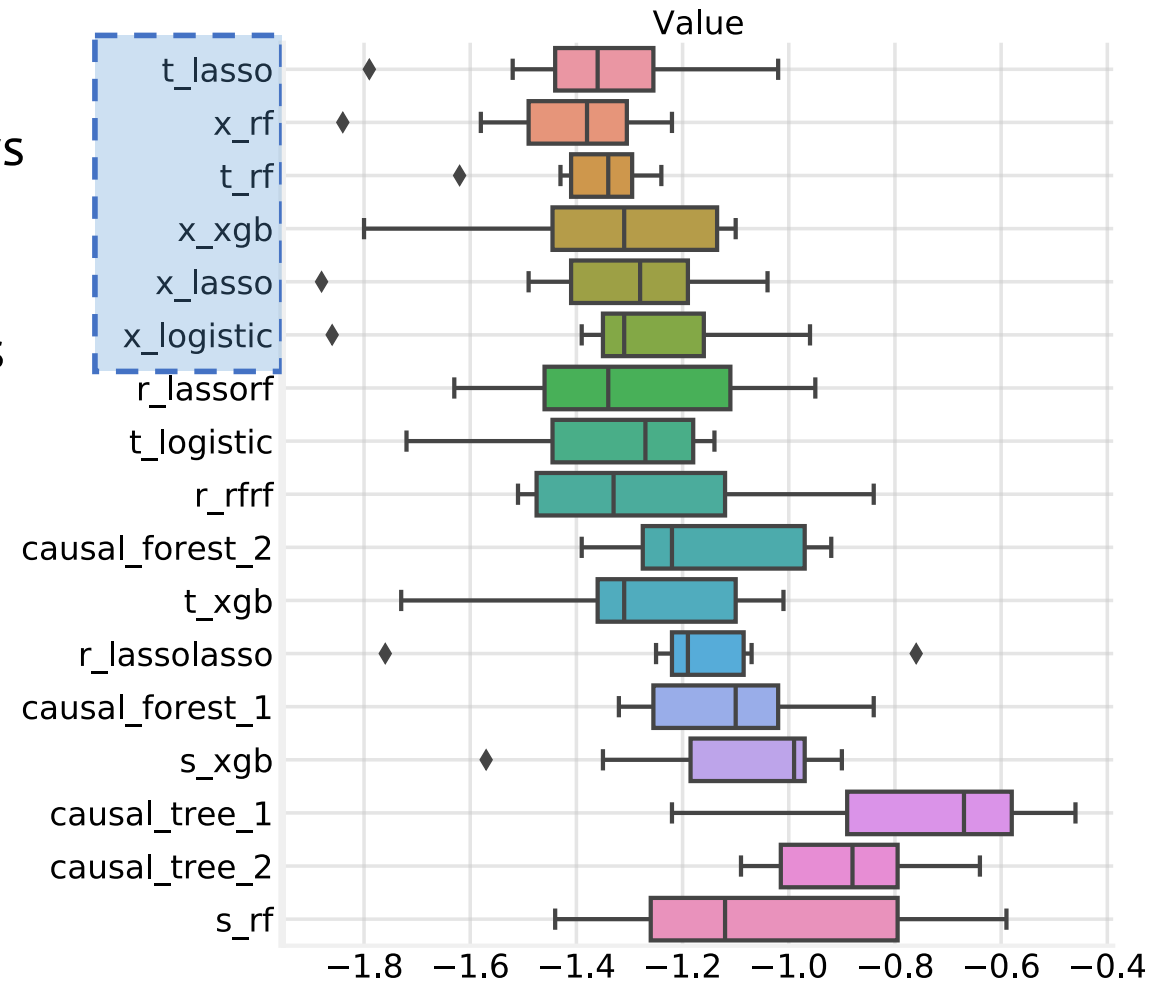
7 data perturbations:

3 from feature engineering tuning

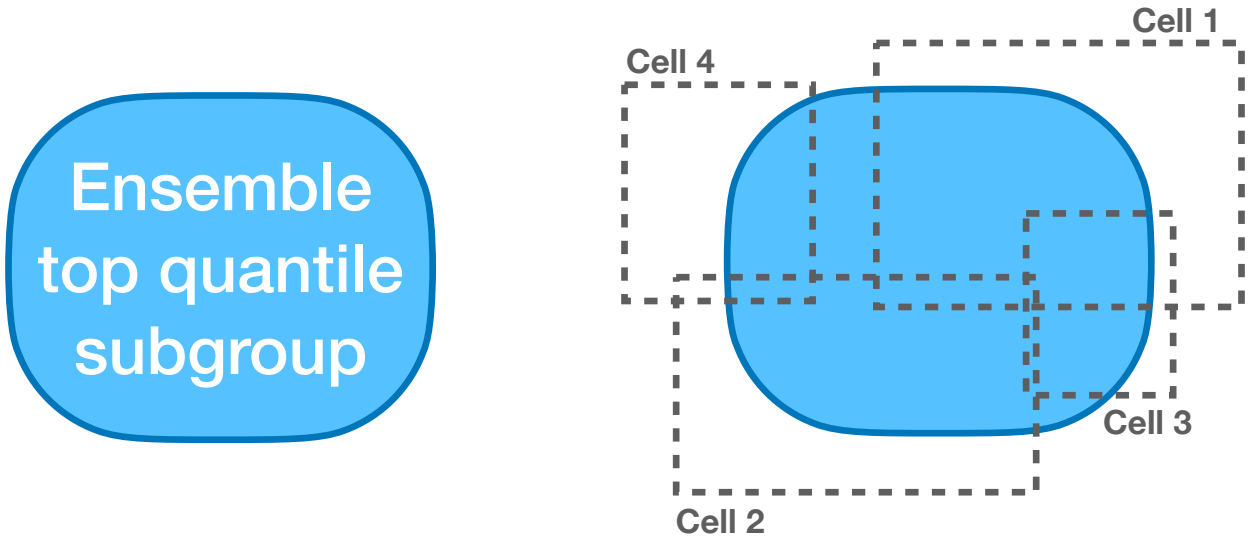
4 from fixed 4 validation: average across the 3 splits on training folds

Prediction screening through ranking the 17 CATE models using t-scores

Selected **6** models always in "top 10" across 7 perturbations



Cell search for interpretable subgroups based on ensemble-6 cate estimate



- A cell / decision rule is a rectangular region defined by constraining the values of some features
- **Objectives -- simplicity:** Few stable disjoint cells---each based on few features & **Coverage**

StaDISC finds interpretable subgroups

Vioxx when compared to **Naproxen**

Disproportionately reduced GI risk for patients with
History of GI event
History of hypertension & Prior usage of steroids
Old age & prior usage of steroids

Disproportionately increased CVT risk for patients with
History of atherosclerosis
Usage of aspirin indicated by FDA
Old age and male gender*

*Poor generalization on test set (no event)

External evaluation

Evaluation results with APPROVe study

Vioxx when compared to **Placebo**

Disproportionately increased GI risk for patients with
✅ History of GI event
History of hypertension & Prior usage of steroids#
Old age & prior usage of steroids#

#Very small subgroups (no event)

Disproportionately increased CVT risk for patients with
✅ History of atherosclerosis
✅ Usage of aspirin indicated by FDA
✅ Old age and male gender

Summary

Veridical data science (trustworthy AI) through

- **PCS** framework (workflow and documentation on github)
- PCS case study and evaluation via staDISC using VIGOR and APPROVe
- **PCS** generates testable results for external validation (experiments or other studies)
- Domain knowledge is imperative in PCS

Data Science Book by Yu and Barter with MIT Press

Free on-line interactive copy (plan: 2021 summer)

Veridical Data Science: A Book

Bin Yu[1,2] and Rebecca Barter[1]
[1]Department of Statistics, UC Berkeley
[2]Department of Electrical Engineering and Computer Science, UC Berkeley


Berkeley
UNIVERSITY OF CALIFORNIA

What skills does the book teach?

Veridical Data Science (VDS) will teach the critical thinking, analytic, human-interaction and communication skills required to effectively formulate problems and find reliable and trustworthy solutions.
VDS explains concepts using visuals and plain English, rather than math and code.
The primary skills taught are:

Critical thinking

Readers will learn to:

Formulate answerable questions using the data available
Scrutinize all analytic decisions and results
Document all analytic decisions
Appropriate common techniques to unfamiliar situations
Deal with real, messy data

Technical skills

Data processing
Data cleaning
Exploratory Data Analysis
Data merging

Algorithmic
Dimension reduction
Clustering
Least Squares & ML
Regularization

Stability-based inference
Inference
Causal Inference
Perturbation Intervals
Trustworthiness Statements

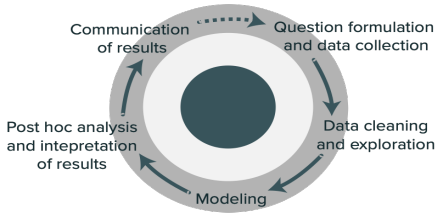
Communication

Exploratory Visual Summaries
Preparing explanatory visual and numeric summaries for explaining data and findings to an external audience

Written reports
Preparing written analytic reports for case studies based on real, messy data

Core guiding principles for the book

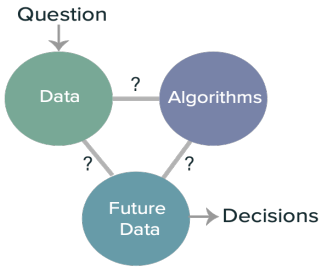
The DS Lifecycle



The Data Science Lifecycle is an iterative process that takes the analyst from problem formulation, data cleaning, exploration, algorithmic analysis, and finally to obtaining a verifiable solution that can be used for future decision-making.

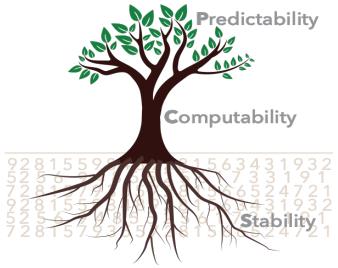
Blending together concepts from statistics, computer science and domain knowledge, the data science life cycle is an iterative process that involves human analysts learning from data and refining their project-specific questions and analytic approach as they learn.

Three realms



Readers will learn to view every data problem through the lens of connecting the three realms:
(1) the question being asked and the data collected (and the reality the data represents)
(2) the algorithms used to represent the data
(3) future data on which these algorithms will be used to guide decision-making.
Guiding the reader to connect the three realms is a means of guiding the reader through the data science lifecycle.

PCS framework



The PCS framework provides concrete techniques for finding evidence for the connections between the three realms.
Predictability: if the patterns found in the original data also appear in withheld or new data, they are said to be predictable. If an analysis or algorithm finds predictable patterns, then these patterns are likely to be capturing real phenomena.
Computability: algorithmic and data efficiency and scalability is essential to ensuring that the results and solutions (e.g. a predictive algorithm) can be efficiently applied to new data.
Stability: minimum requirement for reproducibility. If results change in the presence of minor modifications of the data (e.g. via perturbations) or human analytic decisions, then there might not be a strong connection between the analysis/algorithms and the reality that underlies the data.

Intended Reader/Audience

Anyone who wants to learn the intuition and critical thinking skills to become a data scientist or work with data scientists.
Neither a mathematical nor a coding background is required.
VDS could form the basis of a semester- or multi-semester-long introductory data science university course, either as an upper-division undergraduate or early graduate-level course.

Interested? Get in touch!

Bin Yu
Email: binyu@stat.berkeley.edu
Website: https://www.stat.berkeley.edu/~binyu/Site/Welcome.html

Rebecca Barter
Email: rebeccabarter@berkeley.edu
Website: www.rebeccabarter.com
Twitter: @rlbarter

Berkeley Computing, Data Science, and Society **(CDSS led by Jenifer Chayes)**

About ▾ Academics ▾ Research ▾ News ▾ Events ▾ Support Us

Data Science Major
Data8 (1000+ students)
Data100 (1000+)
Data102 (200+)
co-created and co-taught by stats and EECS faculty

A data science program for everyone

Data8 on EdX

Berkeley UNIVERSITY OF CALIFORNIA

Professional Certificate in
Foundations of Data Science

I'm interested

What you will learn

- How to interpret and communicate data and results using a vast array of real-world examples from different domains
- How to make predictions using machine learning and statistical methods
- Computational thinking and skills, including the Python programming language for analyzing and visualizing data
- How to think critically about data and draw robust conclusions based on incomplete information

Expert instruction
3 skill-building courses

Self-paced
Progress at your own speed

4 months
4 - 6 hours per week

$537.30 ~~$597~~ USD
For the full program experience

Thank you!

1. *Veridical data science* . (Yu and K. Kumbier, 2020, *PNAS*)

2. *Stable discovery of interpretable subgroups via calibration in causal studies.* (R. Dwivedi, Y. Tan, B. Park, M. Wei, K. Horgan, D. Madigan, B. Yu, 2020) *Accepted at International Statistical Review) arXiv:2008.10109 (code also available)*

3 *Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist,* (B. Norgeot, …, A. Butte, 2020, Nature Medicine)